Practical 2: Non-linear Regression

Garyfallos Konstantinoudis

Spring Term 2025

You might need the following packages. If you have already downloaded them, skip the following line:

```
install.packages(c("dplyr", "ggplot2", "patchwork", "splines", "blmeco", "mgcv", "dlnm",
          "plotly"))
```

```
library(dplyr)
library(ggplot2)
library(patchwork)
library(splines)
library(blmeco)
library(mgcv)
library(dlnm)
library(plotly)
```

Part 1: Triceps and subscapular skinfold thicknesses using basis functions

Triceps and subscapular skinfold thicknesses provide an index of body fat and midarm muscle circumference provides a measure of muscle mass. The dataset `triceps.csv` contains the age in years, the intriceps and triceps skinfold thickness in cm. In the first part of this practical we would like to examine the association between age and triceps skinfold thickness.

As a first step we load the data.

Question 1.1 Perform the `head()` function to the dataset to understand the data structure. Fit a regression model (Normal, Poisson or logistic depending on the nature of the main outcome) to examine the linear association between triceps skinfold thickness and age. Provide an interpretation of the result.

Question 1.2 What is the main assumption regarding the shape of the relationship between triceps skinfold thickness and age in question 1.1? Can you check if this assumption is valid?

Question 1.3 Fit 3 models with polynomial basis function of order 2, 3 and 4. Plot the results together and discuss. Perform a series of likelihood ratio test to examine which model fits the data best.

Question 1.4 Select the best performing model of the question 1.3 and plot the flexible fit together with the 95% confidence intervals. Hint: Use the function `predict()` to get the standard error and compute the upper and lower limit of the confidence intervals.

Question 1.5 Now use a Fourier basis function and fit three models with one sin/cos pair and period $P = 25, 50, 100$ years. Plot the results together and discuss.

Part 2: Triceps and subscapular skinfold thicknesses with linear splines

Question 2.1 Use the same dataset as before and now fit 2 linear threshold models (not linear splines!), the first with one threshold at 10 years, and the second with two thresholds one at 10 and one at 40 years. Plot the results together and discuss. How do you interpret the results for the different thresholds? Discuss potential limitations of this approach.

Question 2.2 Now without using the `bs()` function or any kind of splines related packages in R, fit linear splines with one threshold at 10 years. Plot, calculate the 95% confidence intervals, interpret the results and compare with the results of the question 2.1.

Question 2.3 Using the `bs()` function, fit a 4 linear basis spline models in the data with one knot each time at 5, 10, 15 and plot all three fits. Which model is the most appropriate for our data?

Question 2.4 Take the `summary()` of the model from question 2.3 with a knot at 10 years and compare the results with the `lm()` fit of questions 2.2. What do you observe?

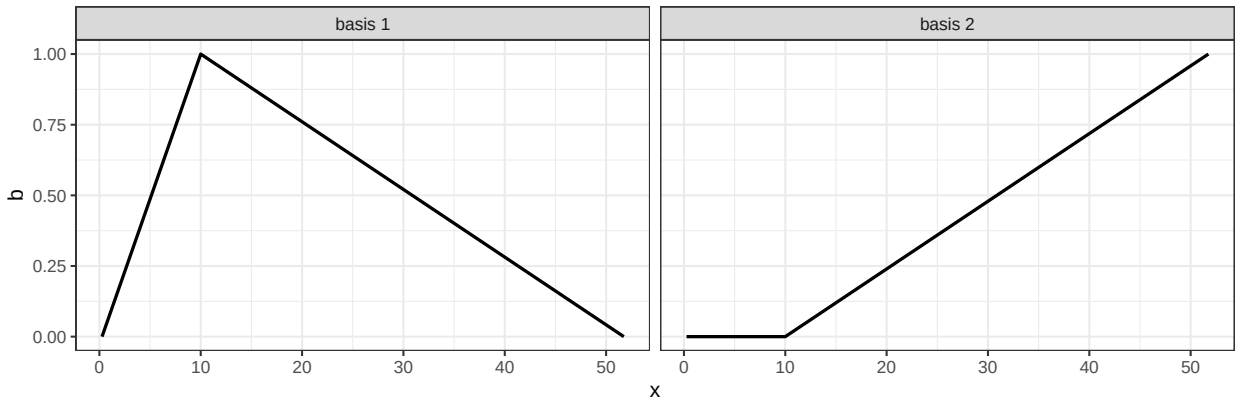
Question 2.5 In this question we will plot the basis function to understand the meaning of the coefficients of the `bs()` fit.

```
b <- bs(triceps$age, degree = 1, knots = 10)
str(b)
```

```
##  'bs' num [1:892, 1:2] 0.951 0.991 0.999 0.964 0.997 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:2] "1" "2"
##  - attr(*, "degree")= int 1
##  - attr(*, "knots")= num 10
##  - attr(*, "Boundary.knots")= num [1:2] 0.26 51.75
##  - attr(*, "intercept")= logi FALSE
```

```
b1 <- b[, 1] ## basis 1
b2 <- b[, 2] ## basis 2

data.frame(x=rep(triceps$age, times=2),
      b=c(b1, b2),
      basis=rep(c("basis 1", "basis 2"), each = triceps$age %>% length())) %>%
 ggplot() + geom_line(aes(x=x, y=b), cex=.8) + facet_grid(cols=vars(basis)) + theme_bw()
```



and now run the following and compare again:

```
lm(triceps$triceps~b1+b2) %>% summary()
```

```
##
## Call:
## lm(formula = triceps$triceps ~ b1 + b2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9714  -1.9977  -0.6263   1.2370  25.3828
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.7936     0.3558  21.906   <2e-16 ***
## b1           -0.4013     0.4800  -0.836    0.403
## b2           10.6424     0.5220  20.388   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.946 on 889 degrees of freedom
## Multiple R-squared:  0.3577, Adjusted R-squared:  0.3563
## F-statistic: 247.6 on 2 and 889 DF,  p-value: < 2.2e-16
```

What do the `bs()` coefficients mean?

to get the estimates of the two separate lines:

```
devtools::install_github("ZheyuanLi/SplinesUtils")
library(SplinesUtils)
RegSplineAsPiecePoly(list_res[[2]], "bs(age, degree = 1, knots = kn[i])", shift = FALSE)
```

How can we interpret the results here? Compare the estimates with the ones retrieved in question 2.2.

Part 3: Triceps and subscapular skinfold thicknesses with splines

Question 3.1 Use the same dataset as before and similar with the question 2.5, use the function `bs()` one knot at 10 and cubic splines

Question 3.2 Use the `lm()` function ad `bs()` to fit cubic splines with one know at 10. What is the interpretation of the `summary()` output?

Question 3.3 Use the function `predict()` and plot the fit of the above model together with 95% confidence intervals.

Question 3.4 Use natural splines instead and plot the fit of the above model together with 95% confidence intervals. Do you observed any differences? Tip: Include boundary knots in the `bs()` function.

Question 3.5 Type `?ns` and check what the `df()` argument does. Now use only the argument `df()` in the `ns()` function and use the values 1, 10, 50, 100. What do you observe?

Question 3.6 Use the function `gam()` from the `mgcv` package and fit a penalized spline. What do you think about the fit?

Part 4: Analysing the HANES dataset

In the fourth part of this practical you will be analyzing the HANES lb dataset. This dataset included information about the age, sex (men=0, women=1) race, location (locode), height, BMI (body mass index), Booze (categorical alcohol consumption), serum calcium (Ser.calc), setum cholesterol (Ser.chol), current smoking, smoking history, number of cigarettes per dat, lifetime pack year and follow up variables including age at death (d.age), year of death (d.year), death from any cause (d.total), death from cancer (d.cancer) and deaths from heart disease (d.heart)

As a first step we load the data and assign `hanes` to a data.frame.

Question 4.1 Fit a logistic regression model for mortality (d.total) with main effects for age, sex, race, booze, smokever, and bp1sys. Interpret the output of the model (focus on the coefficient of the systolic blood pressure).

Question 4.2 Now let's examine the interaction between systolic blood pressure and sex? Is it significant? Interpret the results.

Question 4.3 Repeat that analysis of question 3.1 using the `gam()` function and a smooth function for age. Is age nonlinear?

Question 4.4 Add a smooth function for `ser.chol`. Is it linear?

Question 4.5 Now model d.heart instead of d.total. Is the dependence on `ser.chol` linear?

Part 5: Distribued lag linear models.

In the fifth part of this tutorial, we will use the `chicago` dataset of the `dlnm` package used throughout the lecture and examine the effect of heatwaves.

Question 5.1 What do you think a heatwave is? How should we define it?

Question 5.2 Let a heatwave be the period that we have observed temperatures higher that the 95th percentile of the overall temperature in chicago for at least 2 consecutive days. Create a new variable called heatwave that takes values 0 if it falls within the definition and 1 otherwise.

```
library(dlnm)
chicagoNMMAPS %>% head()
```

```
##         date time year month doy      dow death cvd resp       temp   dptp
## 1 1987-01-01    1 1987     1   1 Thursday   130  65   13 -0.2777778 31.500
## 2 1987-01-02    2 1987     1   2   Friday   150  73   14  0.5555556 29.875
## 3 1987-01-03    3 1987     1   3 Saturday   101  43   11  0.5555556 27.375
## 4 1987-01-04    4 1987     1   4   Sunday   135  72    7 -1.6666667 28.625
## 5 1987-01-05    5 1987     1   5   Monday   126  64   12  0.0000000 28.875
## 6 1987-01-06    6 1987     1   6  Tuesday   130  63   12  4.4444444 35.125
##     rhum     pm10       o3
## 1 95.500 26.95607 4.376079
## 2 88.250       NA 4.929803
## 3 89.500 32.83869 3.751079
## 4 84.500 39.95607 4.292746
## 5 74.500       NA 4.751079
## 6 77.375 40.95607 6.334412
```

You can use the following code to define the heatwave variable:

```
threshold <- quantile(chicagoNMMAPS$temp, probs = 0.95)
chicagoNMMAPS %>%
 dplyr::mutate(
  heatwave = dplyr::case_when(temp >= threshold &
                dplyr::lag(temp, 1) >= threshold &
                dplyr::lag(temp, 2) >= threshold ~ 1,
                TRUE ~ 0)
 ) -> chicagoNMMAPS

chicagoNMMAPS %>%
 dplyr::mutate(
  heatwave = dplyr::case_when(heatwave == 1 |
                 dplyr::lead(heatwave, 1) == 1 |
                 dplyr::lead(heatwave, 2) == 1 ~ 1,
                TRUE ~ 0)
  ) -> chicagoNMMAPS
```

Quantify the effect of heatwave on resp deaths. Interpret the result. Hint: Include heatwave as linear, time as spline, month as spline and adjust for day of week and PM10

Question 5.3 Examine a potential lag effect of the heatwaves on respiratory mortality. Consider lags 1:10 and plot the effect. What do you observe?

Question 5.4 Fit a distributed non-linear model. Consider a linear threshold model with the threshold being the 95th percentile of the temperature and lags 1:10.

```
cb.heatwave <- crossbasis(chicagoNMMAPS$temp, lag=10,
                          argvar=list(fun="thr", thr = threshold),
                          arglag=list(fun="lin"))
```

Use the function `gam()` accounting for time as spline, month as spline and adjust for day of week and PM10. Take the `summary()` of the model and interpret the slopes corresponding to the cross basis function. What is the difference of this estimate with the heatwave estimate?

Question 5.5 Use the function `crosspred()`. Specify `at=-20:30` and `bylag=1`. Recall that `at` gives the temperature values and `bylag` the lags to predict. Retrieve the cumulative relative risk at 30^oC over the lags. What is the interpretation

Question 5.6 Plot a 3D plot to show the lag, temperature and relative risk dimension. Also, plot 2 slices of the 3D plot, one for lag=1 and the second for temperatures=30. If interested, you can use the `plotly` package to produce an interactive 3D plot.